42184 Data Science for Mobility
42577 Introduction to Business Analytics course
Challenge statement

Welcome to this year's challenge! :-)

The topic this year is the *housing market*. It is quite a central one in the lives of almost every person and is responsible for a significant part of the economy of any country. Furthermore, it involves some of the trickiest questions from a machine learning perspective and doesn't require complex computation resources (e.g. no real-time streams, or data from many sensors).

You have access to a dataset from Santiago de Chile, where you have a few details about different house transactions, such as the latitude/longitude coordinates, the transport accessibility context of the area, the characteristics of the household that bought the house.

Project structure

The project has three components:

- *Prediction challenge* - where all groups need to address the same problem (30%)
- *Exploratory component* - where each group is invited to choose their own research question and explore the data accordingly (40%)
- *Report* - Each group should deliver one or more jupyter-notebooks, that should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides the individual comments and reflections (30%)

Figure 1 shows the variables you will have in this dataset. The data is actually provided as an excel sheet that also contains a data dictionary. Notice that some variables (e.g. DirCoordX, DirCoordY) require a lot of treatment in order to be usable.

	0	1	2	3	4
Hogar	100010	100020	100030	100041	100052
paraValidacion	0	0	0	0	0
Sector	7	7	7	7	7
DirCoordX	335180,8019	338410,2114	327863,8248	327864	338480,8152
DirCoordY	6266420,975	6265607,141	6257800,086	6257800	6267296,941
MontoArr	100000	120000	70000	80000	117771
IngresoHogar	450845	1019369	80000	559259	710309
Factor	136,393738	73,843597	180,722809	150,379059	122,001518
AnosEstudio	11	11	10	14	12
CLASE	1	1	1	2	2
Sup_Prom_Constr_Hab_EOD	53,8	59,6	59,5	59,5	43,6
Calid_EOD_norm_inv	0,98	0,98	0,98	0,98	0,98
DensConstr_EOD	0,059	0,033	0,004	0,004	0,086
Dist_est_Metro_MC_KM	23,05171147	21,08017693	34,1478939	34,1478168	19,90879164
Dist_salida_Autop_MC_KM	4,345178548	1,381521077	11,99338943	11,99326807	1,363177127
Tiempo_Com_Stgo	69	84	83	83	94
Ingreso_Promedio_Zona_MM	0,519764709	0,678317	0,408158222	0,408158222	0,498140018
Acc_Comercio_tpte_pub	704,97642	704,97642	704,97642	704,97642	704,97642
Acc_Educacion_tpte_pub	406,0983	406,0983	406,0983	406,0983	406,0983
Acc_Habitacion_tpte_pub	6110,62492	6110,62492	6110,62492	6110,62492	6110,62492
Acc_Industria_tpte_pub	671,08681	671,08681	671,08681	671,08681	671,08681
Acc_Servicios_tpte_pub	719,84272	719,84272	719,84272	719,84272	719,84272
Acc_Comercio_auto	3036,41	3036,41	3036,41	3036,41	3036,41
Acc_Educacion_auto	1781,81	1781,81	1781,81	1781,81	1781,81
Acc_Habitacion_auto	30505,65	30505,65	30505,65	30505,65	30505,65
Acc_Industria_auto	2853,19	2853,19	2853,19	2853,19	2853,19
Acc_Servicios_auto	3058,03	3058,03	3058,03	3058,03	3058,03
CLUSTER7	3	3	3	3	3
CLUSTER2	1	1	1	1	1

Figure 1. Dataframe view

The *prediction challenge* considers the problem of **predicting the type of household that will buy the house, given the area characteristics**. The type of household is given in the variable “CLASE”. Such a model could help a real estate company, to advertise the house to the right segments, for example.

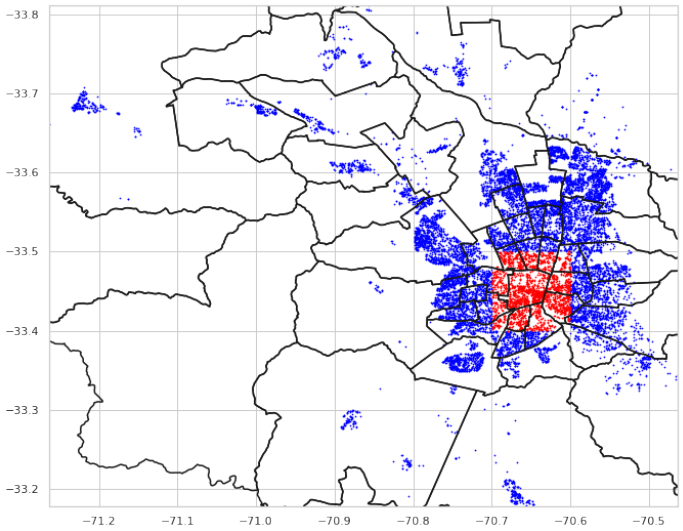

Figure 2. Train set (in blue), test set (in red)

You should test your model in the center of Santiago (see Figure 2), particularly the bounding box with coordinates 70.7E, 33.4S to 70.6E, 33.5S. The train set should be all other available data points (if you want to use a development set, you need to extract it from the train set).

It is very acceptable that you add external relevant data to your model (e.g. statistics from the city, point of interest information). You only cannot “give the answer” in your input data (e.g. since you’re predicting household type, you should not include the household characteristics such as income or education level in the input vector).

In the *exploratory component*, each group needs to address at least one new research question. Some examples are:

- *How is accessibility distributed across space? Are there particular areas that somehow relate to together? Can you explain why?*
- *Can we analyse how the city is distributed in terms of equity (e.g. are there many areas with poor accessibility and low income, and many others with high accessibility and high income)?*

The project will also be positively valued with one or more of the following extensions:

- Extension of the dataset (preferably using Python APIs) with other relevant data (e.g. points of interest from Open Street Map);
- Generation and analysis of insightful visualizations;
- Comparison with other cities, between neighbourhoods, etc.

Evaluation

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions
- Honesty - While it's fine to use others' code (as starting point, these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to always reference the source of that code in you used.

Rules

- Each group should consist of 2-3 students. Exceptions are allowed for single students, but only with strong justification.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the names of the group members (for example, for Pablo, Anders and Mila, it should be Pablo_Anders_Mila.zip).

- In the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defense
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

Important dates

- October 11 – announcement of this challenge statement
- October 25 – Communication of group members (to camara@dtu.dk and rodr@dtu.dk)
- November 15 – Descriptive statistics notebook – this notebook should present preliminary analysis on the dataset, and other datasets obtained by the group, including data preparation, data cleaning, exploratory analysis of patterns and insights from the data. The students are invited to look for relevant questions to the dataset. Submit through CampusNet
- December 6 – Final submission – all materials, including report notebook. Submit through CampusNet